PCT/AU00/00796

#2

4

AU 00/796

Patent Office
Canberra

21 JUL 2000

I, LEANNE MYNOTT, TEAM LEADER EXAMINATION SUPPORT AND SALES hereby certify that annexed is a true copy of the Provisional specification in connection with Application No. PQ 1286 for a patent by TELEFONAKTIEBOLAGET LM ERICSSON filed on 30 June 1999.

COMMONWEALTH OF AUSTRALIA
AUSTRALIA
PATENT OFFICE

WITNESS my hand this
Fourteenth day of July 2000

L. Mynott

LEANNE MYNOTT
<u>TEAM LEADER EXAMINATION</u>
<u>SUPPORT AND SALES</u>

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)

P/00/009
Regulation 3.2

AUSTRALIA
Patents Act 1990

PROVISIONAL SPECIFICATION

FOR THE INVENTION ENTITLED:

"A SCALABLE COMPUTER SYSTEM"

Applicant:

TELEFONAKTIEBOLAGET LM ERICSSON

The invention is described in the following statement:

A SCALABLE COMPUTER SYSTEM

This invention relates to computer systems and more particularly, but not exclusively, to a computer system which is required to have a large number of computer processors for use in large scale system applications.

Recent developments in telecommunications and intelligent networks, primarily involving the field of photonics, are resulting in a rapid expansion of bandwidths available for communication. The available bandwidth is currently growing at a rate involving, roughly, a factor of two every two years, and it is anticipated that communication bandwidths may increase by at least three orders of magnitude over the next ten years.

In order to match this rapid growth in the level of telecommunications, equivalent computing power is required. It is therefore desirable to design a computer based system architecture that is massively scalable. The scalability is essentially driven by the number of independent users, as in mobile phone devices, rather than the complexity or size of an individual application.

Hitherto, there has been no known solution to the problem of designing a massively scalable architecture for the intelligent network (IN) domain. There have been applications, such as encryption, that have been run over large numbers of independent Internet-connected systems, based on the concept that the Internet itself is an extremely large computer system. However, the Internet is a hierarchical system and such applications do not address the problem of designing a massively scalable computer system for use in the intelligent networks. Consistent system architectures that are known in that area are many orders of magnitude less than what will be required in the near future.

It is therefore desirable to provide a massively scalable computer system including a large number of processors in which each processor can communicate effectively with other processors without regard to their locations.

According to one aspect of the invention, there is provided a large scale computer system comprising a multiplicity of nodes, each node having a plurality of interconnected processors, said nodes being arranged in a network with

neighbouring sets of nodes of the network forming neighbourhoods of fully interconnected nodes, wherein random links are provided between nodes of different neighbourhoods in the network whereby each processor of the system can communicate effectively with other processors regardless of their location in the network without full connectivity in the network.

The present invention utilises a message based approach and a "small-world" network architecture in which a relatively small number of random cross-links of nodes or vertices in a network can result in small characteristic path lengths for the transfer of messages between nodes or vertices in the network regardless of their location.

In very large networks with only local connections most vertices or nodes in the network are separated by many links. In regular lattice or ring structures with local connections between vertices or nodes, the characteristic or mean distance of the path length (L) between two vertices or nodes grows approximately linearly with the size of the network. On the other hand, in a network with only random connections between vertices or nodes, the characteristic path length (L) grows only logarithmically with the number of vertices or nodes. However, such a network is a poorly clustered environment in which a sparse number of random connections can result in some vertices or nodes not being connected to other nodes.

A small world network is one which falls between a regular network and a completely random network in that vertices or nodes in local neighbourhoods or clusters are interconnected to each other, and a relatively small number of random links or connections are provided between nodes of different neighbourhoods of the network.

Structural properties of small-world networks can be defined mathematically in terms of their characteristic path length $L(p)$ and clustering coefficient $C(p)$. $L(p)$ measures the typical separation between two vertices in a network (a global property) and $C(p)$ measures the clustering or connections in a typical neighbourhood (a local property), where p, is a probability factor. In a regular network having n nodes and k edges or links per nodes, $p = 0$, $L(0)$ grows linearly

with n, the number of nodes in the network, and C(0) depends on the specific geometry or 'wiring' of the network.

In a completely random network with only random connections, p = 1, L-random grows only logarithmically with n, whereas C-random $\simeq$ k/n $\ll$ 1. Examples of regular, small-world and random networks are shown in Figure 1.

Studies of natural, social and man-made networks, such as the neural network of the worm C. eleganis, a collaboration of film actors and the power grid of the western U.S.A. (Collective dynamics of "small-world networks" by Duncan J. Watts & Steven H. Strogatz, Nature, Vol 393, pp 440-442) have shown that such networks are "small-world" networks and that there is a broad interval of p over which L(p) is almost small as L-random, yet C(p) is much greater than C-random. This is shown in the graph of Figure 2.

Watts and Strogatz have thus demonstrated numerically that a few random global connections are sufficient to turn a regular network into a small-world network to reduce the path length of the number of links for effective communication between nodes or vertices drastically. Hanspeter Herzel in his article entitled "How to Quantify Small-World Networks" (Fractals, Vol 6, No. 4 (1998) 301-303) has also derived a simple formula for the mean connectivity k^{eff} of nodes in a small-world ring network with k edges of links per node, in which a fraction of p connections are re-wired randomly:

$$k^{eff} = k \cdot p$$

When $k^{eff} > 1$, the characteristic path length L(p) may be expressed as follows:

$$L(p) = \frac{\ln(n)}{\ln(k^{eff})} \qquad \frac{\ln(n)}{\ln(k) + \ln(p)}$$

Such a logarithmic curve approximates the curve L(p)/L(0) in Figure 2.

In the present invention, the number of nodes in the network, the number of nodes in the neighbourhood, the number of nodes per neighbourhood and the ~~number of cross-links in the network may each be varied for different applications,~~

provided that the network functions as a small-world network with large clustering C(p) and small average path lengths L(p).

The cross-links in the small-world network of the present invention may be chosen completely at random. Alternatively, a pseudo-random selection process may be used to select the cross-links between neighbourhoods to convert a regular network into a highly clustered small-world network with a relatively small average path length.

In one preferred embodiment of the invention a mean connectivity falling substantially in the range from about 1.5 to about 2.0, and preferably about 1.6, may be achieved by appropriate choice of the number of nodes per neighbourhood, the connectivity of each neighbourhood and the number of cross-links relative to the total number of nodes in the network. By way of example, in a computer system having 50 neighbourhoods of 10 nodes arranged in a ring network with about 50 cross-links between neighbourhoods, each node is connected to 9 other nodes in its neighbourhood (k = 9) and the probability factor p = 50 x 2/500. Thus

$$k^{eff} = 9 \times \frac{100}{500} = 1.8.$$

The number of interconnected processors in each node of the computer system may also vary for different applications. With the development of photonics, it is envisaged that up to about 256 nodes, each containing up to eight processors per node may be fully connected together by optical fibres and a high speed switch to form a single neighbourhood in a small-world network. With a large number of such neighbourhoods in the small-world network connected by a relatively small number of cross-links, it is possible to achieve a massively scalable computer system with a very large number of total processors in the system which are able to communicate with each other in an effective manner.

A preferred embodiment of the present invention will now be described, by way of example only, with reference to the accompanying drawings in which: -

Figure 1 is a diagram showing the transition from a regular ring network to a random network via a "small-world" network;

Figure 2 is a graph showing variations in clustering and average path length with increasing randomness in a ring network.

Figure 3 is a diagram of a single node computer system which has limited scalability;

Figure 4 is a diagram of a split node computer system;

Figure 5 is a diagram of a four node computer system having four processors per node;

Figure 6 is a diagram of a small-world architecture for a computer system in accordance with the invention.

The computer system of Figure 3 is based on an Erlang/Open Telecom Platform (OTP) running on a single node. The computer system 10 includes hardware 12, an operating system 14, a display 16 and a keyboard 18 and a suite of programs 20 which include application programs 22 (eg. in programming languages, Erlang and C), sourced programs 24, run-time programs 26, a library 28, and a database 30. The system may be linked to an external database 32 if required.

The single node provides very good system development facilities, including an Erlang real time environment, or interpretive environment. However, this is achieved at the expense of potential performance owing to the interpreter/operating system layers.

The single node computer system of Figure 3 may be linked to other similar nodes by an asynchronous transfer node (ATM) switch, such as the AXD-301 switch with satisfactory performance. However, this switch has a scalability of 1:30 which is orders of magnitude less than that which is required for ultra high communication bandwidths.

Referring to Figure 4, there is shown a split node computer system in which an OTP node 34 is split into two closely couple nodes: a COTS (commodity of the shelf) system 40 and a multi-processor (MP) Erlang Engine 50. The COTS system is essentially the base system and may comprise of a UNIX operating platform 41, application programs (eg. in C, C++, Java and Erlang) 42, a disc drive 43, graphics 44, an Internet modem 45, an Internet interface (TCP/IP) 46 and an input/output interface (I/O) 47 for communicating with the Erlang Engine 50.

The Erlang Engine 50 is a shared memory MP system running Ericsson core software 52 and possible outsourced software 54 in Erlang on top of an optimised message passing kernel (56) such as QNX. The Erlang Engine 50 also has an I/O interface 58 for communicating with the COTS system 40. One processor of the MP set can be devoted to monitoring software, the remainder to functional processing.

The split node OTP system 34 of Figure 4 may form part of a network which includes a plurality of regional processors (RP) 61 and support processors (SP) 62 for operators. The regional and support processors 61, 62 are connected to central processors CP A 63 and CP B 64 and to each other by a high speed RP bus 65, and the MP Erlang Engine 50 includes a high speed interface 59 for communicating with the central processors CP A 63 and CP B 64.

The split node OTP system 34 can be linked to other computer systems by a switch 70 such as an AXE-10. For this purpose an AXE programming system (APS) 72 may be provided. In a telecommunications application the interface shown to the AXE-10 may be implemented as a high speed Ethernet, primarily due to the availability of the Ethernet PLEX (Programming Language for Exchanges) blocks existing on AXE-10. The Erlang to PLEX interface has been demonstrated in two modes, firstly with the AXE-10 controlling the links, as in a call forwarding application. The second mode is with OTP in control, with an application such as remote changes to AXE-10 tariff tables.

The limit for scalability for the Erlang Engine is for a maximum of eight processors, given that it is a shared memory environment. The eight processors, together with a demonstrated 5 x speed up from the move to compiled code, plus 2 x moving from Unix to QNX gives a scale-up of 80 from the base system.

Referring to Figure 5, there is shown a four node/four processor per node computer system which demonstrates that it is possible to link up several MP Erlang Engines of the type shown in Figure 2 through a high speed blocking switch. Each node of the computer system of Figure 3 is an Erlang Engine 150 with four processors 152 running on a QNX kernel 156. Each Erlang Engine 150 may

A network connected in this manner would normally be termed a regular network in which messages from one node, eg A_1, must pass through a relatively large number of links to reach a node at the opposite side of the network, eg C_4.

A small-world network such as shown in Figure 6, however, differs from a regular network in that a relatively small number of random cross-links are provided between neighbourhoods of nodes. In the example of Figure 4, there are three cross-links CL_1, CL_2 and CL_3. Cross link CL_1 directly connects nodes A_2 and C_1 with each other; cross link CL_2 directly connects nodes B_2 and E_3 with each other and cross-link CL_3 directly connects nodes D_1 and E_4 with each other. It will be apparent from Figure 6 that there is a marked increase in connectivity with only a relatively small number of cross-links between neighbourhoods of nodes in the ring network. For instance, a message to be sent from node A_1 to C_2 can pass along an edge link EL to node A_2, along cross-link CL_1 to C_1 and then along an edge link to C_2, rather than along four loop links to C_1 and an edge link EL to C_2. Also, a message from B_4 to E_2 can pass along a loop link to B_2, along cross-link CL_2 to E_4 and then along another loop link LL to E_2, rather than along five loop links.

It will be seen from Figure 6 that a small-world network is an architecture that can be used to link together a relatively large number of computer processors while retaining effective connectivity between the processors. Even if each neighbourhood is a four node/four processors per node system similar to that of Figure 5, a total of eighty processors can be effectively linked together by the 20 node ring network of Figure 6 with only 3 cross-links.

The small-world network of Figure 6 is, however, only a relatively small scale example of such a network. Within the scope of the present invention it is contemplated that at least 500 neighbourhoods of nodes each having up to 256 fully connected Erlang Engines per node with 8 processors per Erlang Engine could be linked together in a small-world network to achieve a total system of over one million processors, with the small-world network architecture providing effective connectivity between the nodes with only a relatively small number of cross-links, say 50, between the neighbourhoods.

For example, the total number of individual Erlang processes running on such a system will be of the order of 256,000,000 assuming 2000 active processes per Erlang Engine node. This gives around 25.6 million lines of code which is of the order of magnitude already envisaged for large software systems project.

It is envisaged that a massively scalable computer system in accordance with the invention has widespread applications. In the telecommunications field, the scalability could cover applications such as mobile telephone services for stockbroking, betting and other services.

It will be appreciated that various modifications and alterations may be made to the present invention as described above without departing from the scope and spirit of the present invention. For instance, as mentioned above, the size of the network, the neighbourhoods in the network, the number of nodes per neighbourhood, the number of cross-links between neighbourhoods and the number of processors per node may be varied for different applications.

DATED: 30 June 1999

CARTER SMITH & BEADLE
Patent Attorneys for the Applicant:

TELEFONAKTIEBOLAGET LM ERICSSON

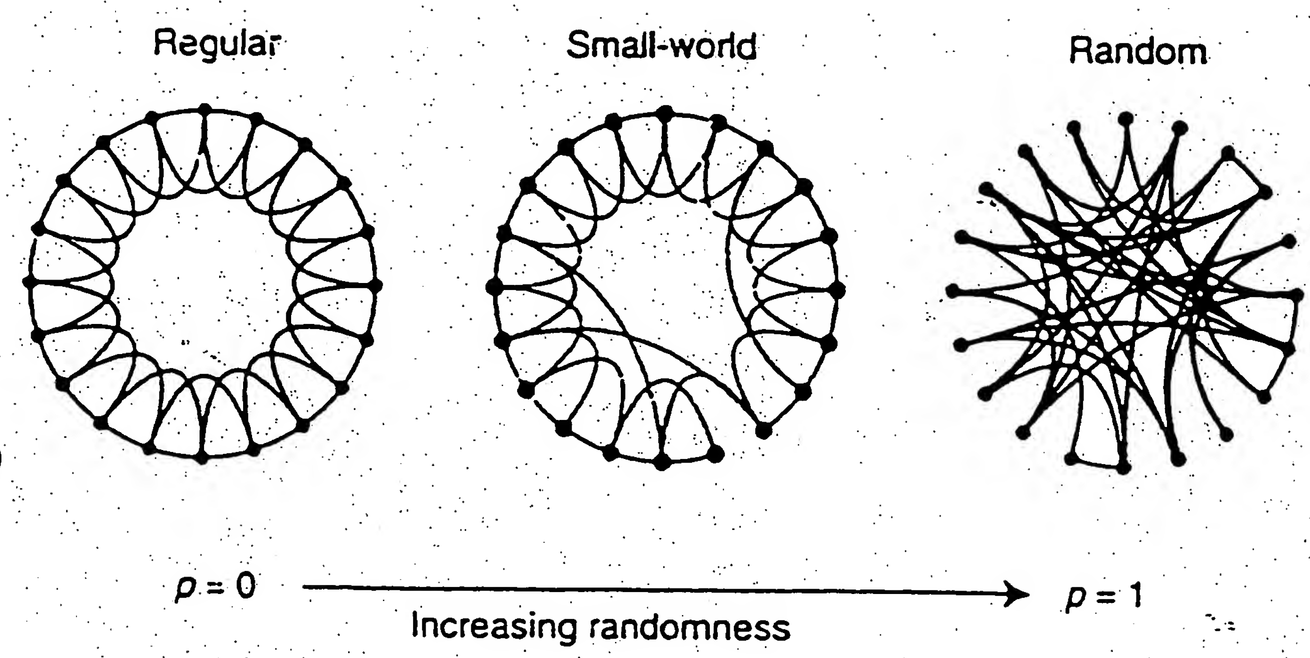

Regular
Small-world
Random
p = 0
Increasing randomness
p = 1

FIG. 1

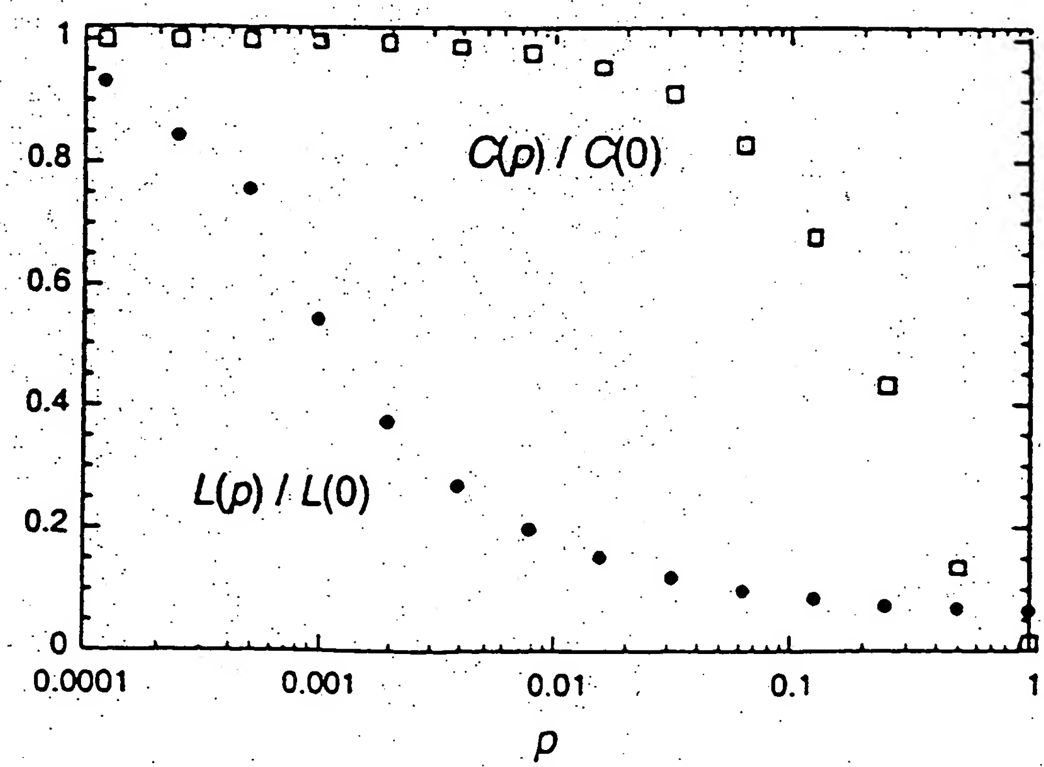

1
0.8
0.6
0.4
0.2
0
0.0001
0.001
0.01
0.1
1
p
C(p) / C(0)
L(p) / L(0)

FIG. 2

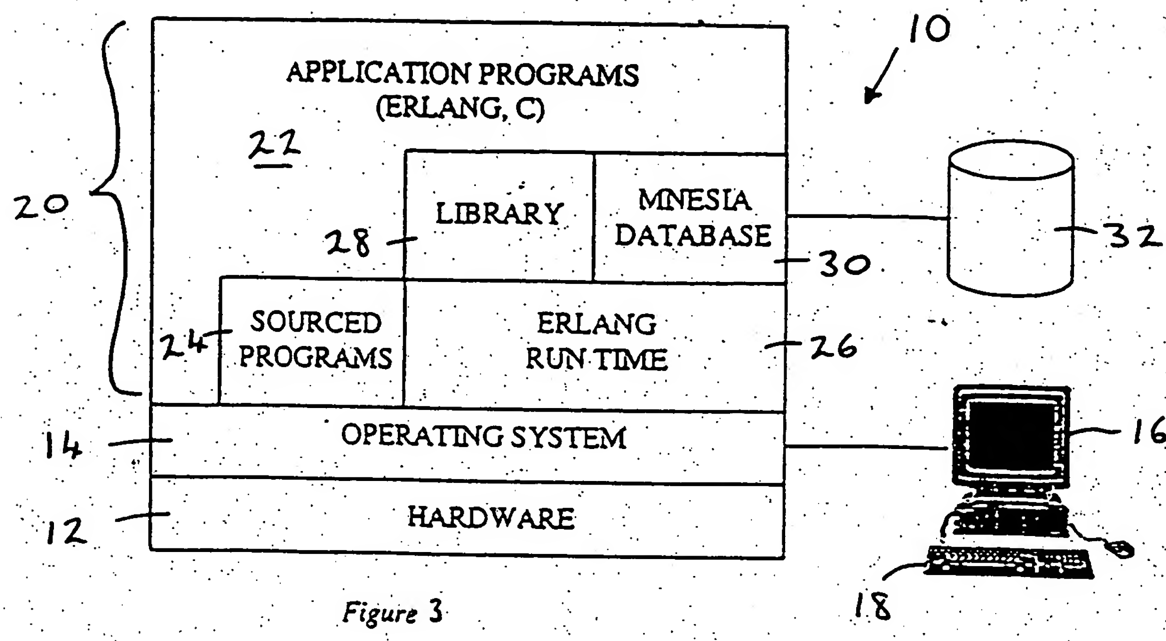


Figure 3

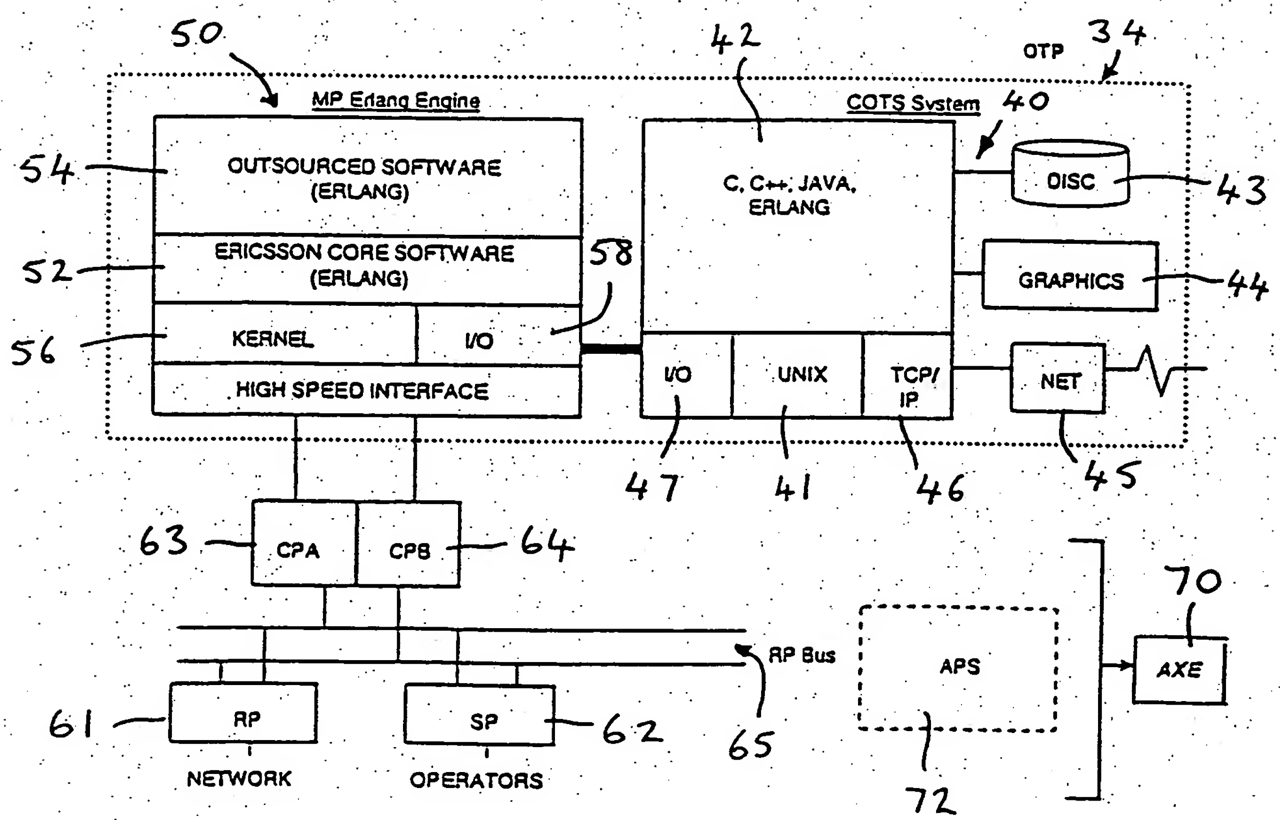
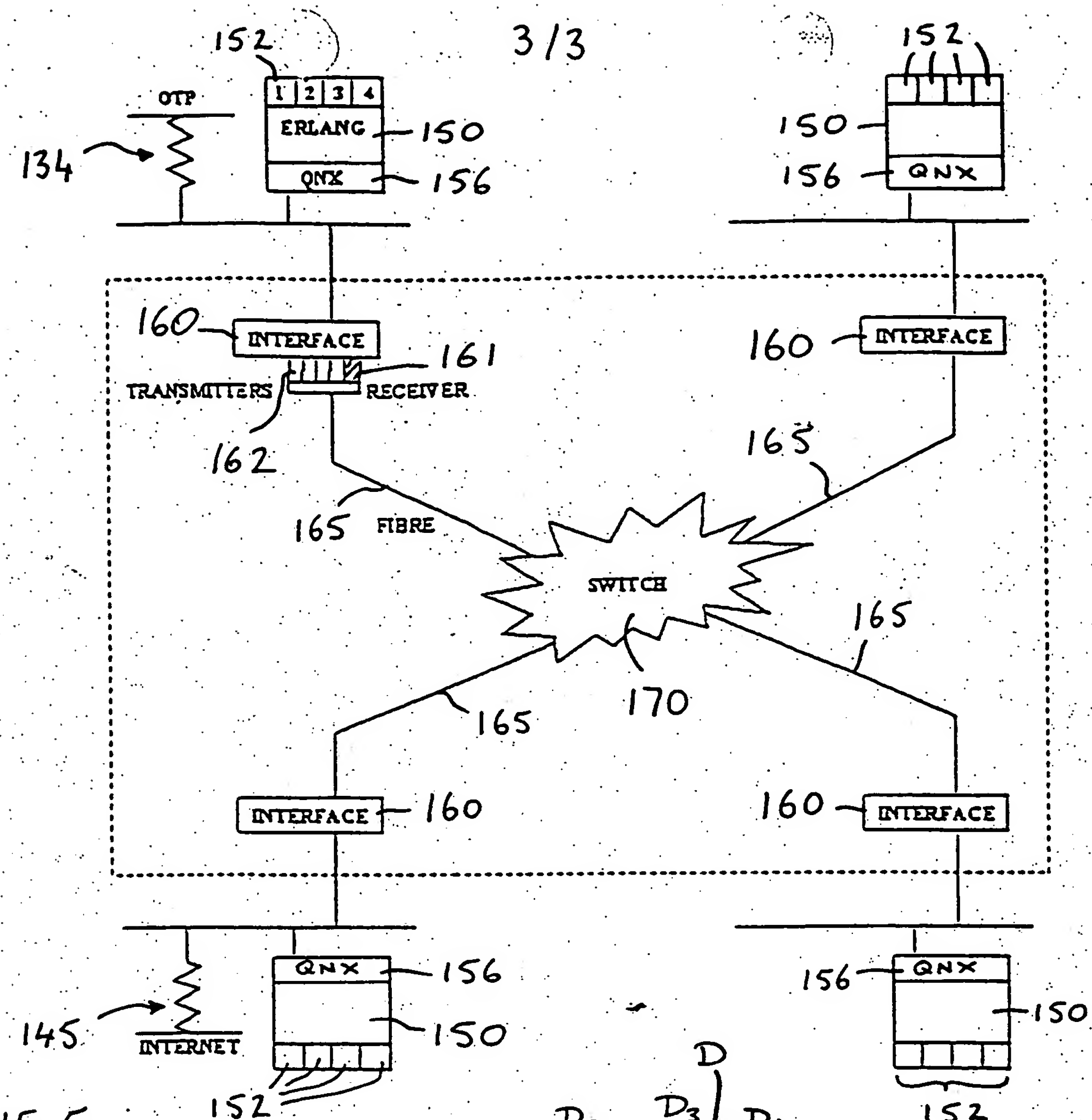
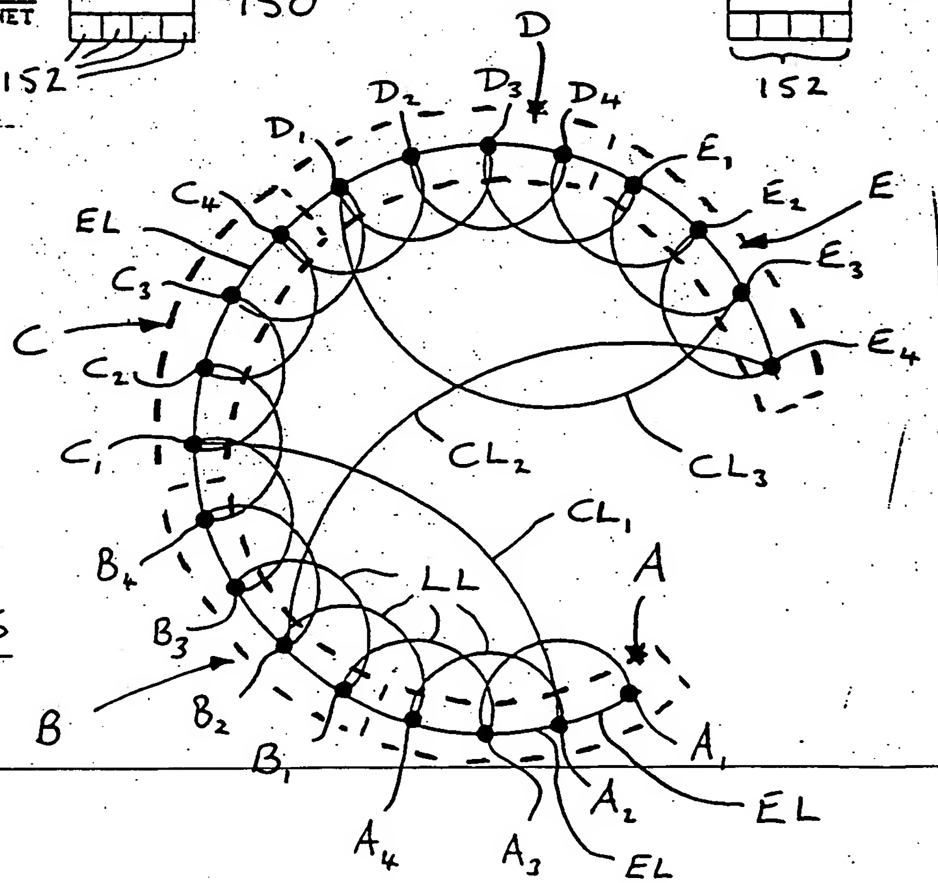


Figure 4



152
1 2 3 4
OTP
134
ERLANG
150
QNX
156
152
150
156
QNX
160
INTERFACE
161
TRANSMITTERS
RECEIVER
162
160
INTERFACE
165
165
FIBRE
SWITCH
165
170
165
INTERFACE
160
160
INTERFACE
QNX
156
145
INTERNET
150
152
156
QNX
150
152

FIG. 5



D
D2
D3
D4
D1
E1
C4
E2
E
EL
E3
C3
C
E4
C2
C1
CL2
CL3
CL1
B4
LL
A
B3
B
B2
A1
B1
A2
EL
A4
A3
EL

FIG. 6